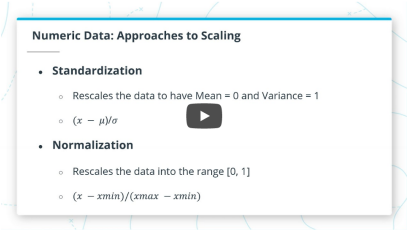
Scaling Data

Scaling data means transforming it so that the values fit within some range or scale, such as 0–100 or 0–1. There are a number of reasons why it is a good idea to scale your data before feeding it into a machine learning algorithm.

Let's consider an example. Imagine you have an image represented as a set of RGB values ranging from 0 to 255. We can scale the range of the values from 0–255 down to a range of 0–1. This scaling process will not affect the algorithm output since every value is scaled in the same way. But it can speed up the training process, because now the algorithm only needs to handle numbers less than or equal to 1.

Two common approaches to scaling data include **standardization** and **normalization**.



Standardization

> ***Standardization*** *rescales data so that it has a mean of 0 and a standard deviation of 1.*

The formula for this is:

> $(x - \mu)/\sigma$

We subtract the mean (μ) from each value (x) and then divide by the standard deviation (σ). To understand why this works, it helps to look at an example. Suppose that we have a sample that contains three data points with the following values:

```
50
100
150
```

The mean of our data would be `100`, while the sample standard deviation would be `50`.

Let's try standardizing each of these data points. The calculations are:

```
(50 - 100)/50 = -50/50 = -1
(100 - 100)/50 = 0/50 = 0
(150 - 100)/50 = 50/50 = 1
```

Thus, our transformed data points are:

```
-1
0
1
```

Again, the result of the standardization is that our data distribution now has a mean of 0 and a standard deviation of 1.

Normalization

> ***Normalization*** *rescales the data into the range [0, 1].*

The formula for this is:

> $(x - xmin)/(xmax - xmin)$

For each individual value, you subtract the minimum value ($xmin$) for that input in the training dataset, and then divide by the range of the values in the training dataset. The range of the values is the difference between the maximum value ($xmax$) and the minimum value ($xmin$).

Let's try working through an example with those same three data points:

```
50
100
150
```

The minimum value ($xmin$) is `50`, while the maximum value ($xmax$) is `150`. The range of the values is $xmax - xmin$ = 150 − 50 = 100.

Plugging everything into the formula, we get:

```
(50 - 50)/100 = 0/100 = 0
(100 - 50)/100 = 50/100 = 0.5
(150 - 50)/100 = 100/100 = 1
```

Thus, our transformed data points are:

```
0
0.5
1
```

Again, the goal was to rescale our data into values ranging from 0 to 1—and as you can see, that's exactly what the formula did.

QUESTION 1 OF 3

Which of the below refers to **standardization** and which refers to **normalization**?

Submit to check your answer choices

DESCRIPTION	STANDARDIZATION OR NORMALIZATION?
Rescales the data to have mean = 0 and standard deviation = 1	Standardization
Rescales the data into the range [0, 1]	Normalization
$(x - xmin)/(xmax - xmin)$	Normalization
$(x - \mu)/\sigma$	Standardization

SUBMIT

Standardize -5,10,15. Knowing that the mean is 7 and the standard deviation is 10. Use commas to separate numbers and keep one decimal place (e.g. 1.0,2.3,3.0 or -1.5,1.1,2.4)

(-1.2,0.3,0.8)

RESET

Normalize -5,10,15. Knowing that the mean is 7 and the standard deviation is 10. Use commas to separate numbers and keep one decimal place (e.g. 1.0,2.3,3.0 or -1.5,1.1,2.4)

(0.0,0.7,1.0)

RESET

NEXT